View email as a webpage | SCINET website

USDA-ARS SCINet Newsletter: April 2021

Contents

- **Announcements**
- **Research Highlights**
- **Atlas Corner**
- **SCINet Training Program**
- **Meet the VRSC**
- **Meet our SCINet Fellows**
- **How to Get Started**
- **Contribute / Contact**

Announcements

New storage allocations on Ceres, Atlas, and the NetAPP:

SCINet is happy to announce that additional storage has been added to the system (the NetAPP). The advantages to you will be that: (1) your data will be more secure, (2) either on Ceres or Atlas, you will have more flexibility to use high speed storage for computational needs, and (3) the new structure will promote clean project folders. In the near future, we will be providing storage on the NetAPP that will allow HPC users (Ceres or Atlas) to store files between analyses. As part of this effort, we are encouraging a community cleanup of files that are either no longer needed, are temporary files that can be recreated as needed, or are duplicate files.

Proposed plan and timeline: For projects with <=1TB currently used space, we encourage files to be cleaned up, but we will not change the storage allocation on the HPCs, and will add 1TB of space on the NetAPP. For projects >= 1TB, the plan is to decrease storage currently being used on Ceres or Atlas and add storage to the NetAPP based on current usage (see

Table below). We will have an open period for comments, concerns, and suggestions. If you would like to have input to this process then send an email to ARS-SCINet-CSIO@usda.gov by **May 1.**

Projects will have until June 30 to clean up files. On July 1, we will reduce storage on Ceres and Atlas for projects >1TB and provide NetApp storage to each project based on the table below.

We hope this process will enable you to work more effectively on the ARS HPCs with this new storage opportunity. We look forward to your comments and suggestions, and to working with you to ensure efficient use of this new storage capacity.

Project folders

Current Ceres or Atlas Storage	New Amount on Ceres or Atlas (starting July 1)	Overall Action
<=1TB	1TB	no change in space
>1 to 6TB	1-3TB	50% decrease in space
>6TB	6TB	X% decrease in space to get down to 6TB

Storage based on Project folder

Current Ceres or Atlas Storage	New NetAPP Storage	Overall Action
<=1TB	1TB	no change
>1TB	75% of present occupied space on Ceres or Atlas	25% decrease

SCINet Corner:

SCINet is launching a new way to answer your questions. VRSC experts are moderating a **monthly** Zoom breakout room mixer. Please come and ask about SCINet accounts, equipment, and resources, as well as how to utilize Ceres and Atlas. SCINet Corner will be a way for people to share knowledge, discuss projects, and identify the resources needed to make progress on their projects. Participants are encouraged to share information on tutorials, conferences, and other upcoming SCINet training opportunities. Together, we can discuss ideas, problems, and form collaborations to address your questions. Come talk science with us!

- Where: Zoom link sent out the Monday before and the night before
- Register: https://forms.gle/7DcBoBvbGcjQDBP38

SCINet Forum:

SCINet has launched a new discussion forum that will be replacing Basecamp over the next month: https://forum.scinet.usda.gov. It should provide better discussion, search, and control of what emails you receive.

- The first time you visit, select "Sign Up" and you will be asked to log in with your existing SCINet account login and two-factor authentication credentials (what you use to log in to Ceres or Atlas). For instance: if your SCINet Password is "Password" and your Verification Code is "123456", then enter "Password123456".
- Once you log in, basic forum usage can be found here. Need help installing or troubleshooting software or programs? Need help with login or account info? These topics and more are now on the SCINet Forum.

We encourage you to take a minute and sign in today!

The AI Innovation Fund grant call is coming soon:

ARS Artificial Intelligence Center of Excellence will be releasing a grant call to support research projects that transfer artificial intelligence and machine learning (AI/ML) methods into agricultural research, or develop a prototype data product that uses AI or ML. We expect to fund 4 to 6 projects of up to $100,000 each. Funds will need to be spent this fiscal year so projects should have a short timeline or involve partnerships that can be funded through agreements. For more information on the grant see the grant announcement. The proposal deadlines will be announced in the coming weeks.

Applications to become a SCINet fellow posted:

There are 15 new opportunities posted through ORISE to become a SCINet fellow--please share with your networks! Four positions require a Master's degree and 11 require a Ph.D. For more information on the positions visit: https://www.zintellect.com/Catalog and enter the keyword: *SCINet*

Research Highlights

Faster, better metagenome analysis: one genome, one contig for metagenome samples sequenced with PacBio HiFi reads

By: Derek Bickhart

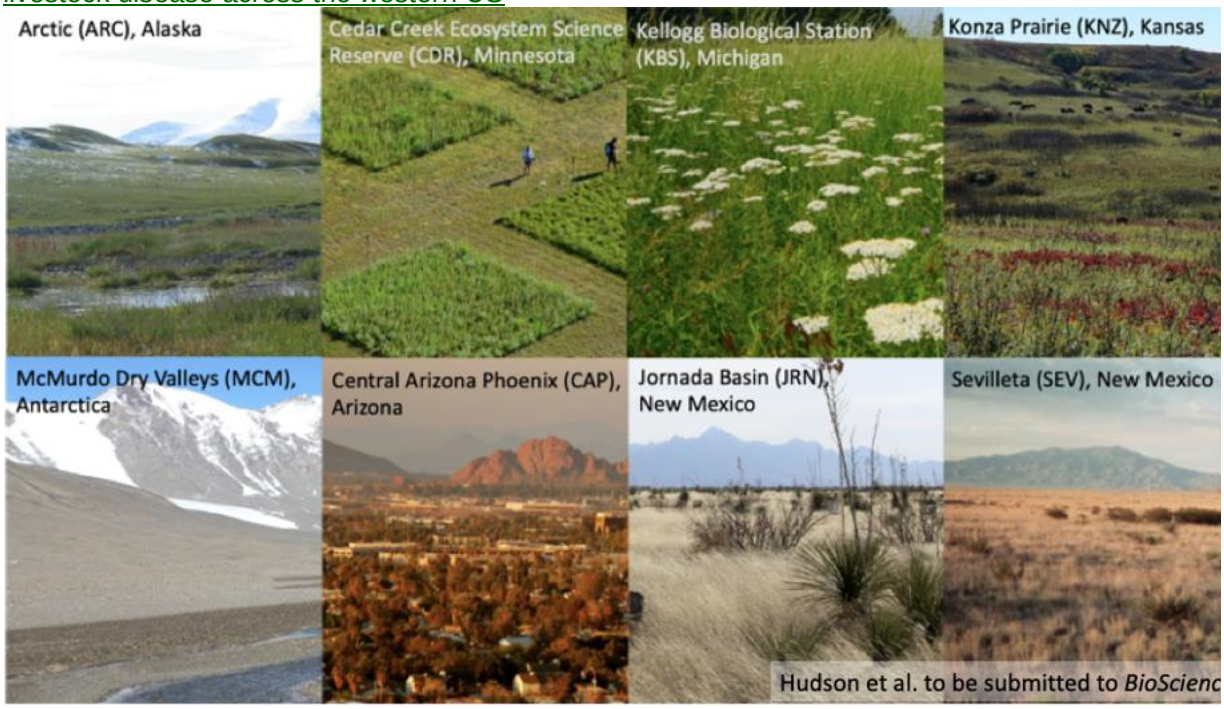
All plants and animals require interactions with microbes to survive and thrive, such as nitrogen-fixing bacteria associated with plant roots and beneficial bacteria in the gastrointestinal tracts of animals. Studying the genomes of all microbes present in a system (metagenomics) helps us to understand how microbial systems impact the health and productivity of plants and animals. However, the large size and complexity of microbial genomes, as well as processes such as horizontal gene transfer that cause microbial genomes to shift constantly, create many challenges in studying agriculturally relevant metagenomes.

Fortunately, recent advances in technology and access to high-performance computing systems make metagenome assembly and characterization much more practical.

The **SCINet Ceres cluster** has been instrumental in our attempts to characterize agricultural metagenomes by providing the computational power necessary for each project. A complicated metagenome can require over 100 billion bases of DNA sequence to be generated in order to identify a large majority of the microbes in the system. These billions of bases of DNA are spread among hundreds of millions of DNA sequence reads, which are akin to puzzle pieces of their original microbial genomes. We can use computer algorithms to recreate genomes from these pieces, but this task is similar to completing a puzzle with hundreds of millions of pieces when we don't know how the end result should look! Using the distributed computing power of the Ceres cluster, we can assemble the final metagenome from one of our samples in just a few days, a task that would take more than a year if we used just one CPU processor core. Access to this computing resource leaves us with more time to analyze the results for interesting biological information.

Read on to learn more about Derek's work here.

Climate driving change: from dryland agro-ecosystems across the globe to the spread of a livestock disease across the western US



By: Amy Hudson*

**Amy is one of 10 postdocs in SCINet's first postdoc cohort. Scroll down to our "Meet our SCINet Fellows" section below for a short introduction to Amy and other featured fellows Ibukun (Timothy) Ayankojo and Lucas Heintzman.*

Rising global temperatures have cascading effects on Earth's agro-ecosystems. To better

understand and contextualize these effects, we leveraged the long records of observational and experimental data collected across a suite of sites in the Long-Term Ecological Research (LTER) Network, including two Long-Term Agroecosystem Research sites. Debra Peters and I organized a large team of researchers from eight dryland sites across a wide geographic and production gradient (shown by images above). Drylands receive low amounts of rainfall, and plants at these sites are limited in how much they can grow by the amount of water available. Warming amplifies water limitations at these sites, and increases the frequency and intensity of wildfires, dust, and flooding events. We highlight and compare the importance of these multiple drivers to primary production with feedbacks to global climate in an upcoming issue in BioScience special. Kellogg Biological Station and Jornada Experimental Range are two sites from this study that are also part of the 18-site Long-Term Agroecosystem Research (LTAR) Network.

Climate can also drive the spatial spread of livestock disease, such as Vesicular Stomatitis (VS), that I am studying with the VSV Grand Challenge group, whose multi-disciplinary members include ARS scientists Debra Peters, Luis Rodriguez, Lee Cohnstaedt, Barbara Drolet, Justin Derner, and Emile Elias along with our collaborator from USDA APHIS, Angela Pelzel-McCluskey. The virus, VSV, enters the US from Mexico every few years and then spreads to cover much of the western US with a tendency towards regions that are cooler and wetter the following year. These results are consistent with the characteristics of habitat conditions and life cycles of the black flies and biting midges insect vectors for this disease. Large-scale modes of climate variability, such as the El Niño Southern Oscillation (ENSO), can drive changes in regional climate to influence the location and timing of VSV across the Mexico border. The SCINet HPC resources allow us to use a large number of datasets (>400) in AI frameworks towards better prediction of the spread of VSV as a model for predictive disease ecology across the US.

Do you use SCINet for your research?
Contact SCINet-Newsletter@usda.gov for a chance to be featured in the newsletter!

Atlas Corner

Photo of Senior computer specialist Joey Jones at the Atlas machine

SCINet's HPC systems and support are expanding with the recent addition of Atlas, a new HPC maintained by Mississippi State University. Here are some key points for users to consider when deciding to implement their workflow on Atlas. More documentation on Atlas can be found here: https://www.hpc.msstate.edu/computing/atlas/

Processor: Atlas is a machine with differing processors and memory configurations. Your code may run faster with additional cores per node (48 cores per node).

Network: Atlas' interconnect utilizes Mellanox's HDR technology. This provides 200Gbps switch connections with 100Gbps node communications. This can be beneficial for distributed (multi-node) jobs dependent on fast/high-bandwidth communication.

Storage: Atlas' project storage space is different from Ceres, so data must be moved in or copied out. Quotas are independent and quota increases should be requested through the VRSC. Users can move data in different ways such as: ssh enabled transfer methods such as scp and rsync to any of Atlas' DTNs (Data Transfer Node); slurm jobs using the 'service' partition; Globus--Atlas' endpoint name is "msuhpc2#Atlas-dtn".
Programs: A large, and growing, assortment of applications are available. Atlas has many development tools including compilers, MPI implementations, and containers. Atlas uses a hierarchical module structure to modify one's environment. The commands "module avail" and "module –r spider" can be used to list available programs. Applications that depend on specific compilers, mpis or singularities can be made available once the dependency is loaded. For example, to list all the singularity containers, execute "module load singularity" then "module avail". Additional software packages and container images may be available by contacting the VRSC.

How do I get an Atlas account? If you already have a Ceres account, your **Atlas account has been automatically created**. Atlas' login credentials and dual factor authentication are shared with Ceres. If you do not have a Ceres account, contact the VRSC to request one.

SCINet Training Program

SCINet-funded Training

- In March 2021, Mississippi State University held three UAS (unmanned aircraft systems) workshops for USDA-ARS scientists that included **Developing a UAS Program, UAS: Missions, Operations, and Planning, and Data Wrangling**. Over 75 ARS scientists and university collaborators completed the trainings designed to assist them in utilizing UAS in their research. These workshops will be offered again later this spring and in the coming summer months. To learn more about these trainings, read below and reach out to Dixie Cartwright.
    - **Developing a UAS Program** assists researchers in developing a robust UAS program, from planning a successful UAS research program to aircraft registration and certifications. This workshop is ideal for those trying to decide if UAS is appropriate for their research.
    - **UAS: Missions, Operations, and Planning** is designed for research pilots of UAS to help them take to the skies quickly, safely, and with a sound approach to gathering research data. The workshop provides an overview of vehicle

types, regulations, operations planning, flight skills and currency requirements, and reporting.

    - **Data Wrangling** provides data analysts an overview of the various common UAS payloads, such as RGB and multispectral imagers, hyperspectral sensors, and LIDAR. In this workshop, sample processing flows are demonstrated in two of the popular commercial photogrammetry suites, Agisoft Metashape and Pix4D Pix4Dmapper, producing a reflectance-corrected and georeferenced orthomosaic from each software.

- The SCINet Arthropod Genomics Research (AGR) working group is currently hosting the international AGSX Virtual Symposium Spring 2021. This series consists of four monthly sessions from February through May 2021, facilitated by AGR lead Brad Coates and with technical expertise from Glenn Hanes. AGR members and external collaborators are organizing and leading the sessions: "Insect Genomic Technologies to Improve Food Applications" by Brenda Oppert (AGR), "Arthropod Genomics and Genome Engineering" by Lindsey Perkin (AGR), "Application of New Genomic Tools and Techniques in Arthropods" by Marce Lorenzen (North Carolina State University), and the "Honeybee Workshop" by Sonia Eynard and Alain Vignal (INRAE, France). There are currently 351 registered participants from 45 countries, and session videos are being posted to the i5K YouTube channel. This online symposium provides the exchange of ideas and tools among a diverse group of international scientists, including those from ARS, who are involved in genomics research during the COVID-19 pandemic.
- **The Carpentries R and Python Workshops:** Multi-day online software and data carpentry workshops co-sponsored by SCINet and The Carpentries are being offered in March/April/May 2021. March workshop topics included Unix, Git, and R or Python. April and May workshop topics will cover data organization and cleaning, and Intro to Python or R. All workshops are full, but you can join the priority list for the next round of Carpentries workshops by contacting Kathy Yeater.
    - **The Carpentries instructor training** is available to qualified applicants who desire to become certified Carpentries workshop instructors. This training teaches participants The Carpentries pedagogy and provides the resources to instruct Carpentries workshops in Unix, git, SQL, OpenRefine, R, Python, and more. Contact Kathy Yeater for more information about Instructor Training.
- **Coursera.org Certified Courses:** The SCINet Initiative and the AI Center of Excellence are excited to provide training opportunities through Coursera. A limited number of Coursera licenses are available to ARS scientists and support staff to complete training focused on scientific computing and artificial intelligence. These licenses rotate every quarter to maximize participation by ARS scientists and staff. Successful completion of courses and specializations will result in widely recognized certificates and credentials. The application deadline for the next round of licenses has been extended to April 23. Please visit the Coursera Training page of the SCINet website for more details about the program and application process.

Free Online Computational Training (Self-paced):

Make use of your work-from-home time with computational training! A large list of free tutorials and courses has been compiled on the Free Online Training page. Training topic areas include Python, R, SAS, and MATLAB programming; statistics; data science concepts; AI and machine learning; GIS; Google Earth Engine; Git and GitHub; reproducibility, productivity, and

integration management tools; and bioinformatics and ecology domain learning. Know of additional free training opportunities? Send them to SCINet-Newsletter@usda.gov.

SCINet Online Science Tutorials:

Browse our growing set of SCINet science tutorials created by ARS scientists and the SCINet Virtual Research Support Core. Our ARS Science Tutorials page includes Ceres Onboarding and Intro to Unix for new HPC users, two geospatial computing tutorials, a QTL Analysis tutorial for sequencing in R, and machine learning training material.

Meet the VRSC

The SCINet Virtual Research Support Core (VRSC) manages the day-to-day operations and maintenance of the SCINet HPC systems and is also involved with the networking aspects of SCINet. The VRSC is staffed by talented IT administrators and engineers with expertise in high-performance computing system and network engineering, operations, maintenance, and administration. Additionally, the VRSC provides research IT support, computational expertise in various research domains, and support for SCINet-funded workshops/training for SCINet users.

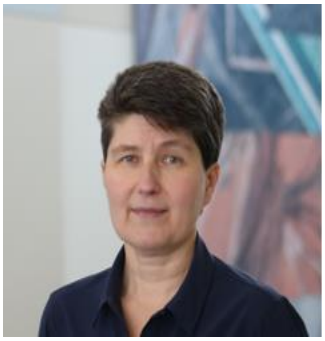
In this issue we're introducing some of the VRSC personnel who facilitate the Ceres HPC system at Iowa State University (ISU) in Ames, Iowa: **Jim Coyle, Marina Kraeva, David Orman,** and **Andrew Severin**. In the next issue, we'll introduce the VRSC at Mississippi State University who support Atlas.

Jim Coyle

Jim Coyle is the Director of High Performance Computing at ISU. His team manages Ceres and three additional research computing and teaching clusters at ISU.
In the first four years of his team running the ISU clusters, research productivity for those involved in HPC at ISU increased 47%. Faculty and students were able to concentrate on their research rather than on running machines, and they were able to run larger studies on the common hardware. Jim enjoys conversations with researchers, advocating for increased computational resources for ISU faculty, and promoting computational collaborations.

Following an undergraduate in Physics, Jim earned a Ph.D. in Mathematics, developing algorithms for efficiently and accurately solving scientific problems. During his career, he has worked as a consultant to ISU faculty and computer and software companies on improved reliability and performance of computer software.

Marina Kraeva



Marina Kraeva is a Senior Research Computing Systems Analyst at ISU with over 25 years of experience in HPC. She received her Ph.D. in Computer Sciences from the Novosibirsk State Technical University, Russia in 1999 and soon moved to Ames, Iowa to join the HPC group at ISU. In the last 20 years, she has been providing HPC support to ISU faculty and students, developing error-detection tools, maintaining HPC clusters and managing HPC projects. Marina enjoys helping users and solving problems. When the ISU HPC group was selected to support SCINet's Ceres cluster, she concentrated on customer interaction, onboarding new users, writing documentation, and helping scientists run their jobs on the cluster. She is glad that most of the information can now be found in one place on the SCINet website.

David Orman

David Orman is a Senior Research Computing Analyst at ISU. He works with a team to manage Ceres and three additional research and teaching clusters at ISU. On the Ceres project David works with assorted automation tools to setup and maintain the clusters and projects. He is also heavily involved with AWS account setup and provisioning, Geneious, Jira, Confluence, GA, as well as producing training materials and videos for ISU HPC and Ceres. David has worked at ISU for 26 years and has BSs in Aerospace Engineering and Computer Science, and an MS in Computer Engineering.

Andrew Severin

Andrew Severin manages the Genome Informatics Facility at ISU. He leads a team of experts tasked with enabling USDA scientists to translate big data into informative data on their specific scientific questions. His team has done this by coordinating workshops, meeting with working groups and developing training resources in the form of tutorials and workbooks. His academic background was in biochemistry with a Ph.D. in NMR spectroscopy. His transition into bioinformatics did not occur until his postdoc in 2009 with Randy Shoemaker who studied soybean genetics with the USDA in Ames, Iowa. Andrew enjoys solving problems and brainstorming with researchers. He encourages researchers to check out two workbooks that are actively being developed; https://bioinformaticsworkbook.org and https://geospatial.101workbook.org. Contributions of tutorials and suggestions are welcome.

Meet our SCINet Fellows

Here are introductions to a few of SCINet's first cohort of postdoctoral fellows. SCINet fellows are tasked with developing collaborative research projects that span multiple ARS units and lead to SCINet working groups, as well as working on their own research projects that utilize the SCINet high-performance computing resources (Ceres and Atlas). They also contribute to non-research projects that further the SCINet Computing Initiative, such as the SCINet website, newsletter, and computational trainings. Check out past newsletters for introductions

to Jennifer Chang, Yanghui Kang, Shawn Taylor, Alicia Foxx, Kerrie Geil, and Melanie Kammerer. We're now introducing: **Ibukun (Timothy) Ayankojo, Lucas Heintzman, and Amy Hudson**.

Ibukun (Timothy) Ayankojo, Soil Scientist

Timothy started his SCINet postdoc position in August 2020, working with Dr. Alison Thompson at USDA-ARS in Maricopa, Arizona. He completed his Ph.D. program in soil and water sciences from the University of Florida. His research involves a multidisciplinary approach that combines principles in soil sciences, agronomy, ecology, and modeling in providing solutions to fundamental challenges facing crop production in the area of crop nutrition and water management. As a SCINet postdoc, he is currently working on high-performance computing and prediction of geospatial dynamics in cotton, soybean, and sorghum. In this capacity, he is working on the application of computer programming language and artificial intelligence for high-throughput plant phenotyping to identify and evaluate important crop traits associated with improved productivity. This work will result in novel image processing pipelines capable of returning desired phenotypes within seconds or minutes. Timothy is part of the Unmanned Aerial Systems (UAS) for Agriculture Research Working Group developed to tackle research topics associated with the use of geospatial data analysis in agricultural research.

Lucas Heintzman, Ecologist

Lucas is a landscape ecologist and has been a SCINet postdoc with Dr. Debra Peters' research group in Las Cruces, NM since January 2020. He completed his Ph.D. in Biology at Texas Tech University, where he examined anthropogenic drivers of land cover change and associated connectivity dynamics of Great Plains wetlands. The primary focus of his SCINet position is to integrate remote sensing, geographic information systems, and time-series analyses with machine learning to assess and predict dryland vegetation dynamics across multiple scales. Lucas is a member of the SCINet Geospatial Working group and has contributed to the SCINet software and geospatial common data library mission. He is also a member of the nascent Unmanned Aerial Systems (UAS) for Agriculture Research Working Group and hopes to foster collaborations among this group and the wider SCINet community.

Amy Hudson, Climate Ecologist

Amy researches how climate interacts with ecosystems at regional to global scales. Amy began her SCINet postdoc in May 2020 working with Dr. Debra Peters in Las Cruces, NM after recently completing a Ph.D. in Natural Resources from the University of Arizona where she examined the influence of the Northern Hemisphere jet stream on plant growth, phenology, and monarch butterfly migration. Amy is currently leading a cross-site synthesis of long-term ecology research on the impacts of climate change on dryland sites and is involved in several projects examining the influence of broad-scale climate on the spatial spread of the insect-borne virus Vesicular Stomatitis. Amy is the editor of the quarterly SCINet newsletter, serves on theSCINet software and geospatial common data library committee and the SCINet Geospatial Working Group, and has helped with previous carpentries trainings and to organize and facilitate the late-summer 2020 Geospatial workshop.

How to Get Started

Simply request a SCINet account (eAuthentication required) to get started. Upon approval, you will receive instructions for logging into SCINet and accessing Basecamp.
Check out the new SCINet website for more info on how SCINet can enable your research.
Read the SCINet FAQs covering general info, accounts/login, software, storage, data transfer, support/policy/O&M, parallel computing, and technical issues.

Contribute / Contact

For questions about this newsletter, to contribute content, feedback on the SCINet website, or SCINet policy and development questions please email SCINet-Newsletter@usda.gov.

SCINet Newsletter Editor: Amy Hudson

For technical assistance with your SCINet account, please email scinet_vrsc@usda.gov.

SCINet Leadership Team

Deb Peters, Acting Chief Science Information Officer

Rob Butler, Acting SCINet Project Manager

Adam Rivers, Science Advisory Committee (SAC) Chair

Brian Scheffler and Stan Kosecki, Ex Officio

SCINet Website | Comments